UNITED STATES PATENT APPLICATION FOR GRANT OF LETTERS PATENT

Sophie Vrzic
Mo-Han Fong
Hang Zhang
INVENTORS

PRIORITY SCHEDULER

Withrow & Terranova, P.L.L.C.
P.O. Box 1287
Cary, NC 27512
(919) 654-4520

Docket No. 7000-114

PRIORITY SCHEDULER

Field of the Invention

[0001] The present invention relates to wireless communications, and in particular to scheduling data for transmission from a base station to one or more mobile terminals.

Background of the Invention

[0002] Wireless communication networks that allocate communication resources, such as time or frequency, require a scheduler to select data to be transmitted. When multiple users are vying for these resources, the scheduler must analyze the incoming data and determine the data having the highest priority for transmission. Priority has traditionally been based on maximizing overall system throughput or maintaining a certain Quality of Service (QoS) level to ensure that data is transmitted in a timely fashion. When maximizing throughput, users having better channel conditions are favored over those with worse channel conditions. Thus, the users with the less favorable channel conditions are always given lower priority unless time-sensitive data is discovered.

[0003] In wireless systems with adaptive modulation and coding that require fast scheduling, a flexible scheduler is needed that can guarantee each user's QoS while, at the same time, provide some degree of fairness among the different classes of users. In addition to the fairness and QoS requirements, the scheduler should also be able to maximize the system throughput by taking advantage of the different rates that are assigned to the different users.

[0004] System capacity is a function of outage criteria, which is defined as the probability that the QoS on a per-user basis is not satisfied. Wireless-Internet services are characterized in general into two categories: delay-sensitive services and non-delay-sensitive services. For delay-sensitive services, delivery of each packet before a maximum specified delay is critical in order to guarantee acceptable QoS. When the delay bound is exceeded, the packet is considered dropped. The delay requirement can range from tens of milliseconds for interactive services to several seconds for streaming

services. In certain systems, it is recommended that the probability of a packet drop due to the delay bound being exceeded be less than two percent. For non-delay-sensitive services, guaranteeing a maximum delay for each packet is not necessary; however, end user perceived performance is still important.

[0005] There are many problems with existing schedulers in terms of supporting multi-media wireless-internet services. The schedulers are not designed for multi-carrier operation, which makes them unsuitable for multiple carrier – data and voice (MC-DV) environments. Many schedulers prioritize packets based on carrier-to-interference (C/I) ratios. Such schedulers maximize throughput without regard to fairness or minimum throughput requirements and typically schedule delivery for users that are closest to the base station. Schedulers attempting to provide proportional fairness attempt to maximize throughput while at the same time provide some degree of fairness; however, these schedulers are not designed to satisfy the delay requirements of the delay-sensitive users. Another problem with existing proportional fairness schedulers is that they cannot control the degree of fairness. Further, the schedulers fail to address the outage criteria, guarantee minimum data rates, or minimize drop rates for the delay-sensitive users.

[0006] Accordingly, there is a need for a scheduling technique capable of 1) guaranteeing the per-packet delay bound for delay-sensitive services, 2) guaranteeing a required minimum data rate for a particular service or user, and 3) providing adaptive fairness control. There is a further need for a technique with these capabilities that can optimize multi-carrier and multi-user diversity in order to maximize overall system throughput.

Summary of the Invention

[0007] The present invention provides for a scheduling data for transmission by an access point, such as a base station. The scheduling provides adaptive fairness control, which depends on how close the users are to a minimum data rate requirement. If desired, more emphasis can be placed on fairness when there are users close to the minimum data rate requirement and more emphasis on maximizing throughput when all of the users are far from the required minimum data rate. Scheduling can also

guarantee a maximum drop rate for delay-sensitive data, assuming sufficient resources are available, as well as guarantee a minimum data transfer rate for all users by ensuring that users below their minimum requirement have a higher priority than users that exceed their minimum requirement. If there are not enough resources to satisfy each user's minimum data rate due to a failure of the call admission process, then the variance in throughput can be minimized for each class of users. The scheduling can also optimize scheduling parameters for multi-carrier systems by using the number of carriers to determine scheduling parameters for the delay-sensitive users in order to maximize throughput.

[0008] Those skilled in the art will appreciate the scope of the present invention and realize additional aspects thereof after reading the following detailed description of the preferred embodiments in association with the accompanying drawing figures.

Brief Description of the Drawing Figures

[0009] The accompanying drawing figures incorporated in and forming a part of this specification illustrate several aspects of the invention, and together with the description serve to explain the principles of the invention.

[0010] Figure 1 is a block representation of a wireless communication environment according to one embodiment of the present invention.

[0011] Figure 2 is a flow diagram according to one embodiment of the present invention.

Detailed Description of the Preferred Embodiments

[0012] The embodiments set forth below represent the necessary information to enable those skilled in the art to practice the invention and illustrate the best mode of practicing the invention. Upon reading the following description in light of the accompanying drawing figures, those skilled in the art will understand the concepts of the invention and will recognize applications of these concepts not particularly addressed herein. It should be understood that these concepts and applications fall within the scope of the disclosure and the accompanying claims.

[0013] With reference to Figure 1, wireless networks use access points, such as base stations 10, to facilitate communications with access terminals, such as mobile terminals 12, within a select coverage area, or cell. Respective groups of base stations 10 are supported by a communication network 14, which may include mobile switching centers, a public switched telephone network (PSTN), a packet-switched network, or a combination thereof. The communication network 14 is used to transport packets to and from the base station 10. The packets may be communicated in a direct packet-switched manner or on top of a circuit-switched platform. The manner in which the packets are communicated to the base station 10 is not critical to the invention.

[0014] During downlink communications from the base station 10 to select mobile terminals 12, the base station 10 must determine the manner and order in which to transmit the data received in the packets from the communication network 14 to the mobile terminals 12. In multiple carrier systems, the base station 10 will also determine the carrier, or channel, on which to deliver the packets. Accordingly, the base station 10 will include a control system 16 having control plane 18 controlling the flow of data through a data plane 20. For communicating with the mobile terminals 12, the data plane 20 will process packets received from the communication network 14 via a network interface 22 under the control of the control plane 18. The packets are processed into units, which are delivered to radio frequency (RF) transceiver circuitry 24 for transmission. For the sake of clarity, the term "packet" refers to packetized data, which is received by the base station 10 from the communication network 14. The term "unit" refers to packetized data that is transmitted from the base station 10 to the mobile terminals 12. A unit may include all or any part of one or more packets. Although units may directly correspond to packets, units are preferably a given size wherein packets may vary in size from one packet to another. The units may include voice, video, or traditional data.

[0015] The forward link from the base station 10 to the mobile terminal 12 will include one or more channels, which are divided into defined time slots. The RF transceiver circuitry 24 is configured to modulate a given unit as dictated by the control plane 18 and transmit the modulated unit via one or

more antennas 26 during a single time slot. The RF transceiver circuitry 24 is preferably configured to implement different modulation and coding techniques and speeds based on channel conditions, the capabilities of the mobile terminals 12, or required transmission standards. As noted, the RF transceiver circuitry 24 may transmit units over a number of carriers, or channels. Those skilled in the art will recognize the various possible modulation techniques and that multiple units may be transmitted in a given time slot.

[0016] The control plane 18 includes a scheduler 28, which is configured to prioritize and control the delivery order of units to the mobile terminals 12 based on parameters detailed further below. During operation, packets for any number of mobile terminals 12 are received and stored in a buffer 30 associated with the data plane 20. The buffer 30 is segregated into multiple queues, each associated with a given mobile terminal 12. If the packets do not directly correspond to units, the incoming packets are processed into the desired units. The units are stored in the respective queues in the order in which they are received. Preferably, the queues use a first-in-first-out (FIFO) configuration.

[0017] In operation, the scheduler 28 can provide a guaranteed minimum drop rate for delay-sensitive data, assuming sufficient resources are available. The Quality of Service (QoS) target for delay-sensitive users is measured in terms of the percentage of units that are dropped due to the delay bound being exceeded. For real-time services such as streaming video, there is a requirement that no more than two percent of each user's packets be dropped. When enough resources are available, the scheduler 28 can guarantee that no units are dropped. Furthermore, this QoS target can be guaranteed in such a way that it minimizes the impact on throughput. The QoS target is guaranteed by ensuring that units near their individual delay bound have a higher priority than all units from non-real-time services. The start time at which these units have a higher priority is optimized in order to guarantee the drop rate and to maximize throughput. The start time is calculated based on the number of delay-sensitive users, the maximum number of slots needed for transmission, the maximum number of transmission attempts, and the number of carriers. The relative priority

among the time-sensitive units is inversely proportional to how close they are to their delay bound.

[0018] The scheduler 28 can also guarantee a minimum data transfer rate for all users by ensuring that users below their minimum requirement have a higher priority than users that exceed their minimum requirement. If there are not enough resources to satisfy each user's minimum data rate due to a failure of the call admission process, then the variance in throughput can be minimized for each class of users.

[0019] The scheduler 28 can provide adaptive fairness control, which depends on how close the users are to the minimum data rate requirement. This allows for more emphasis to be placed on fairness when there are users close to the minimum data rate requirement and more emphasis on maximizing throughput when all of the users are far from the required minimum data rate. The scheduler 28 can also optimize scheduling parameters for multi-carrier systems, using the number of carriers to determine scheduling parameters for the delay-sensitive users in order to maximize throughput.

[0020] In general, the scheduler assigns a throughput and fairness priority and a delay priority to each unit in the queue. The throughput and fairness priority controls the degree of fairness and ensures the minimum bit rate requirement, while the delay priority ensures that the delay bound is satisfied. The packet with the highest sum of the two priorities is then scheduled for transmission to the required user, on one or more of the selected carriers. In one embodiment of the present invention, the scheduler 28 is configured to control scheduling for multiple channels and facilitates scheduling based on:

- channel conditions, such as a carrier-to-interference (C/I) value for each of the carriers,
- the average data throughput for each user,
- the minimum required throughput for user,
- the maximum queuing delay for each of the delay-sensitive users, and
- the number of carriers.

The basic flow for scheduling is illustrated in Figure 2. Channel conditions 32, such as the C/I ratio value, for each user and each carrier, $\left(\frac{C}{I}\right)_i^{(k)}(t)$, are used to maximize throughput, while the average data rate 34, $\bar{r}_i(t)$, and the required minimum data rate 36, r_i^*, are used to control fairness and to guarantee the minimum data rate for each user. Accordingly, these factors are processed to determine a throughput/fairness control factor (step 100). The maximum queuing delay 38, τ_i, and the number of carriers 40, N_c, are used to guarantee the delay bound for each user in delay-sensitive transfers and are processed to determine a delay bound factor (step102).

[0021] A channel condition represents the quality of the transmission channel from the base station 10 to the mobile terminals 12. The throughput rates may be a function of actual data throughput, channel conditions, or a combination thereof. Channel conditions may vary continuously and may be determined using any number of techniques. For example, (C/I) ratios, which represent a measure of signal power to interference power, may be fed back to the base station 10 from the mobile terminals 12. The scheduler 28 can continuously track channel conditions as well as a current channel condition for each mobile terminal 12. Similarly, the scheduler 28 can keep track of an average and current rate of data throughput for each of the mobile terminals 12. The delay bound typically defines the time in which a unit or series of units must be delivered. Scheduling may be optimized to account for the units' delay bounds and the amount of data to transmit.

[0022] In general, the scheduler 28 assigns a priority to each unit in the queue using the equation

$$P_{ij}(t_n) = F_i(t_n) + D_{ij}(t_n),$$

where P_{ij} is the priority for packet *j* of application *i*. In the above equation, the variable $F_i(t_n)$ corresponds to the throughput and fairness requirements, while the variable $D_{ij}(t_n)$ corresponds to the delay requirement. The throughput/fairness component $F_i(t_n)$ ensures that the required minimum bit

rate is satisfied for each application while, at the same time, providing control for maximizing the overall system throughput.

[0023] The throughput and fairness component $F_i(t_n)$ includes two components. The first component maximizes throughput by giving a higher priority to a user with a higher selected data rate, while the second component guarantees the minimum throughput and controls the degree of fairness by comparing the average throughput to the minimum required throughput. The equation can be written as

$$F_i(t_n) = f(r_i(t_n)) \cdot g(\bar{r}_i(t_n)),$$

where

$$f(r_i(t_n)) = \left[\frac{r_i(t_n)}{r_{\max}}\right]$$

and

$$g(\bar{r}_i(t_n)) = \begin{cases} 2 - \dfrac{\bar{r}_i(t_n)}{\bar{r}_i^*}, & \text{if } \bar{r}_i(t_n) \le \bar{r}_i^* \\ a^{(\bar{r}_i(t_n) - \bar{r}_i^*)/\bar{r}_i^*}, & \text{otherwise} \end{cases}.$$

In the above equations, $r_i(t_n)$ is the selected data rate at time t_n, $\bar{r}_i(t_n)$ is the average data rate at time t_n, $\bar{r}_i^*$ is the required minimum throughput rate and $r_{\max}$ is the maximum possible data rate that can be selected.

[0024] The value for the parameter *a*, used in the equation for *g*, depends on how close the users are to their individual required minimum throughput rate. The parameter *a* can be represented by

$$a = h\left(\min_i \{\bar{r}_i(t_n) - \bar{r}_i^*\}\right),$$

where the minimum is taken over all the users in the sector and *h* is an increasing function. If there is at least one user with a data rate less than the required minimum, then the value for *a* is set to zero.

[0025] The fairness and throughput component $F_i(t_n)$ of scheduler 28 can have the following properties:

-throughput increases as the parameter *a* increases,

-the degree of fairness increases as the parameter *a* decreases,

-if *a* is equal to one then the scheduler is equivalent to a maximum C/I scheduler,

-each user's minimum data rate is guaranteed if there are enough resources,

-if there are not enough resources to satisfy each user's minimum data rate then the variance in the throughput will be minimized for each class of users and if

$\bar{r}_1 = k\bar{r}_1^*$

then

$\bar{r}_2 = k\bar{r}_2^*$; and

-the equation $F_i(t_n)$ is bounded, $0 \le F_i(t_n) \le F_{\max}$.

[0026] Accordingly, the scheduler 28 can guarantee each user's minimum required data rate, while providing emphasis on either fairness or on maximizing throughput. The degree of fairness is set adaptively, and depends on how close the users are to their individual minimum data rate requirements. If there are not enough resources to satisfy each user's minimum data rate requirement, the variance in throughput will be minimized for each class of users.

[0027] The delay component $D_{ij}(t_n)$ of the scheduler 28 guarantees the delay bound by assigning a start time to each delay-sensitive unit, which represents the first time the unit will be given a higher priority than any of the non-delay sensitive units. Among the delay-sensitive units with a scheduling time larger than its start time, the scheduler 28 gives a higher priority to those units closer to its delay bound. The delay component $D_{ij}(t_n)$ of the priority equation can be represented by the following equation

$$D_{ij}(t_n) = \begin{cases} \dfrac{(t_n - t_{\min,ij})}{t_{\max,ij} - t_{\min,ij}} + F_{\max}, & if\ t_n \ge t_{\min,ij} \\ 0, & otherwise \end{cases},$$

where $t_{\min,ij}$ is the first time slot for which user *i* has a higher priority than any of the non-delay sensitive users; $t_{\max,ij}$ is the latest time that the j^{th} unit for user *i* can be sent; and $F_{\max}$ is the maximum value of the fairness priority equation. The value for $t_{\min,ij}$ can depend on the number of slots needed to successfully transmit the unit, the total number of delay-sensitive users in the system, the maximum number of transmission attempts, and the number of carriers. It can be calculated using the following equation:

$$t_{\min,ij} = t_{\max,ij} - \frac{1}{N_c}\sum_{k=1}^{N_d}\left(N_{s,k}\cdot N_{\max}\cdot E\left[N_{p,k}\right]\right) - M .$$

[0028] The term $N_{s,k}$ is the maximum number of slots that are needed to transmit a unit from user *k*, $N_{\max}$ is the maximum number of transmission attempts, $E\left[N_{p,k}\right]$ is the expected number of packets in the queue for user *k*, N_d is the number of delay sensitive users and N_c is the number of carriers. The parameter M is provided to control the maximum drop rate (step 104) and can initially be set to zero. The parameter M can then be either increased or decreased in order to control the drop rate to the required maximum.

[0029] For a multi-carrier system, in order to maximize the system throughput while guaranteeing the required minimum data rate, the priority equation, P_{ij}, is evaluated for each carrier (step 106). Next, the first unit to be transmitted is preferably selected by maximizing the priority across all carriers (step 108). Prioritization can be represented by the following equation:

$$\max_{p,k}\left\{P_p^{(k)}\right\},$$

where $P_p^{(k)}$ is the priority for unit *p* on carrier *k*.

Once selected, the unit having the highest priority is transmitted using the appropriate carrier giving rise to the priority rating (step 110). Then, the next

unit is selected by maximizing the priority equation across the remaining carriers and units. This is given by

$$\max_{\substack{p=p_i \\ k \neq c_i}} \left\{ P_p^{(k)} \right\}.$$

This process continues until units have been scheduled on all the carriers and transmitted accordingly.

[0030] In essence, the novel scheduler 28 of the present invention can provide a guaranteed minimum drop rate for delay-sensitive data, assuming sufficient resources are available, as well as guarantee a minimum data transfer rate for all users by ensuring that users below their minimum requirement have a higher priority than users that exceed their minimum requirement. If there are not enough resources to satisfy each user's minimum data rate due to a failure of the call admission process, then the variance in throughput can be minimized for each class of users. The scheduler 28 can also provide adaptive fairness control, which depends on how close the users are to the minimum data rate requirement. If desired, more emphasis can be placed on fairness when there are users close to the minimum data rate requirement and more emphasis placed on maximizing throughput when all of the users are far from the required minimum data rate. The scheduler can also optimize scheduling parameters for multi-carrier systems by using the number of carriers to determine scheduling parameters for the delay-sensitive users in order to maximize throughput.

[0031] These aspects of the invention can be implemented using alternative equations and relationships than those described in detail above. Those skilled in the art will recognize improvements and modifications to the preferred embodiments of the present invention. All such improvements and modifications are considered within the scope of the concepts disclosed herein and the claims that follow.